IN THE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF MASSACHUSETTS

SINGULAR COMPUTING LLC, Plaintiff, v. GOOGLE LLC, Defendant.	C.A. No. 1:19-cv-12551-FDS Hon. F. Dennis Saylor IV

DEFENDANT GOOGLE LLC'S MEMORANDUM OF LAW IN SUPPORT OF ITS MOTION TO STRIKE EXPERT REPORT OF SUNIL KHATRI, PH.D.

I. INTRODUCTION

In his expert report, Singular's infringement expert Dr. Sunil Khatri relies on a completely new infringement theory, one that Singular did not previously disclose at any time—not in its preliminary infringement contentions in September 2020, not in its supplemental contentions in August 2022, and not by any other means. Singular offers no justification for replacing its prior infringement contentions with this fresh theory, nor could it, because its new infringement theory relies on information that it had before the close of fact discovery in September 2021, and well before its August 2022 motion to supplement its infringement contentions. Google therefore moves to strike the portions of Dr. Khatri's report that, in violation of the Local Rules and the Court's Scheduling Order, rely on a different infringement theory not disclosed in any infringement contentions.

Dr. Khatri opines that the accused Google Tensor Processing Units ("TPUs") meet the claim limitation of a low-precision high dynamic range ("LPHDR") execution unit because they perform a "two-stage" "LPHDR multiplication operation" that is in fact two separate operations

performed by two separate components of the TPUs—rounding performed by what Dr. Khatri calls "precision-reducer circuits" in the Vector Unit ("VPU") and multiplication performed by the multipliers in the Matrix Multiply Unit ("MXU"). But Singular's infringement contentions did not map the "LPHDR execution unit" element to the so-called "precision-reducer circuits" in the VPU, as Dr. Khatri now does. Instead, Singular's infringement contentions relied *solely* on multipliers in the MXU as satisfying the claim limitation of an LPHDR execution unit.

Singular's about face is not surprising: Dr. Khatri now concedes what Google had asserted in response to Singular's infringement contentions—namely that the MXU multipliers perform only exact mathematical operations. Thus, those multipliers cannot meet the claimed error rates for an LPHDR execution unit.

Faced with this admitted flaw in Singular's disclosed infringement theory, Dr. Khatri manufactures a new one to replace it. Specifically, he mashes together one operation in the VPU with a separate operation in the MXU to create what he calls a "two-stage" "LPHDR multiplication operation." Those separate operations take place in physically separate circuitry in different parts of the chip: a rounding operation in the VPU and a multiplication operation in the MXU.

Singular's belated attempt to assert a new theory of infringement through its expert is contrary to the Local Rules, the Court's Scheduling Order, and case law. Moreover, Singular has not even tried to show good cause for its switch to a new theory. The Court accordingly should reject Singular's shifting-sands approach to infringement, hold Singular to its disclosed infringement contentions, and strike Dr. Khatri's new two-stage "LPHDR multiplication operation" infringement theory from his expert report.

II. PROCEDURAL BACKGROUND

Singular served its infringement contentions on September 4, 2020. Singular did not amend its infringement contentions prior to the close of fact discovery and at no time has Singular sought

leave to replace the infringement theory that it disclosed in its contentions. Yet Singular had all the information Dr. Khatri relies on, including Google witness testimony, documents, and source code, before fact discovery closed in September 2021.

Following *inter partes* review ("IPR") proceedings and the Court's July 27, 2022 Order on Claim Construction, Singular sought leave to supplement its contentions in August 2022. *See* Dkt. No. 355. Singular represented that it was "not seeking to add any new theory of infringement." Dkt. No. 355-1 at 1. Instead, Singular asserted that its motion to supplement sought only "to cite to the source code used in the accused TPU v2 and TPU v3 devices." *Id.* at 5. Consistent with those statements, Singular did not seek to add anything to its contentions other than source code citations, and did not attempt to establish good cause to assert a new infringement theory. On October 19, 2022, the Court granted Singular's motion to supplement its infringement contentions to cite to source code. *See* Dkt. No. 372.

On December 22, 2022, Singular served the expert report of Dr. Khatri. Ex. 1.

III. LEGAL STANDARD

The Local Rules provide that a party may supplement its "preliminary patent-related disclosures," including its contentions, "only by leave of court upon a timely showing of good cause." L.R. 16.6(d)(5). The Court's August 4, 2020 Scheduling Order reiterates this requirement, and provides that, except for amendments made up to 30 days before the *Markman* hearing, infringement contentions may be supplemented or amended "only by leave of court, for good cause shown." Dkt. No. 70 at 1. Infringement disclosures serve an important function for the fair and orderly progress of patent litigation by requiring "'parties to crystallize their theories of the case early in litigation and to adhere to those theories once they have been disclosed.'" *Philips N. Am. LLC v. Fitbit LLC*, No. 19-11586-FDS, 2021 WL 5417103, at *3 (D. Mass. Nov. 19, 2021) (citing *O2 Micro Int'l Ltd. v. Monolithic Power Sys., Inc.*, 467 F.3d 1355, 1366 n.12 (Fed. Cir. 2006)).

Courts "may impose any 'just' sanction" in response to a party's failure to comply with patent local rules or a court's scheduling order, "including 'refusing to allow the disobedient party to support or oppose designated claims or defenses, or prohibiting that party from introducing designated matters in evidence.'" *O2 Micro Int'l*, 467 F.3d at 1363 (citing Fed. R. Civ. P. 16(f); Fed. R. Civ. P. 37(b)). Thus, courts routinely strike—and the Federal Circuit has affirmed the striking of—expert reports that espouse an infringement theory not reflected in that party's contentions. *See, e.g.*, *Corus Realty Holdings, Inc. v. Zillow Grp., Inc.*, 860 F. App'x 728, 734-35 (Fed. Cir. June 29, 2021) (affirming district court's striking of expert report asserting previously undisclosed theories of infringement); *Phigenix, Inc. v. Genentech, Inc.*, 783 F. App'x 1014, 1017 (Fed. Cir. Sept. 5, 2019) (same); *Inline Plastics Corp. v. Lacerta Grp., Inc.*, No. 18-11631-TSH, Dkt. No. 124 (D. Mass. Feb. 23, 2021) (granting motion to strike expert report introducing new infringement theory not disclosed as required under Local Rules); *ViaTech Tech., Inc. v. Microsoft Corp.*, No. 17-570-RGA, 2021 WL 663057 (D. Del. Feb. 19, 2021) (same); *KlausTech, Inc. v. Google LLC*, No. 10-cv-05899-JSW (DMR), 2018 WL 5109383, at *3 (N.D. Cal. Sept. 14, 2018) (noting that "expert reports cannot go beyond the bounds of the disclosed infringement theories and introduce new theories not disclosed in the contentions"), *aff'd*, 792 F. App'x 954 (Fed. Cir. Feb. 12, 2020).

IV. ARGUMENT

Singular accuses Google's TPU v2 and v3 of infringement. Ex. 1 ¶¶ 27, 69. The TPUs each have two "Tensor Cores." *Id.* ¶¶ 78, 83. Each Tensor Core has several distinct components, including a VPU, MXU (one in the v2 and two in the v3), and a core sequencer. *Id.* ¶¶ 79, 83. The VPU includes, among other things, a rounding circuit that can round a 32-bit floating point ("FP32") value to a 16-bit floating point ("BF16") value. Ex. 2 (Phelps Depo.) at 55:22-56:9, 92:22-93:1, 94:6-9. Although Google's witness and Singular's counsel both referred to this circuit

as a "rounding circuit," Dr. Khatri instead uses a different name of his own invention: a "precision-reducer" circuit. Ex. 1 ¶ 130 (citing Phelps Depo. at 92). The MXU multiplies a pair of 128 x 128 matrices. *Id.* ¶ 79.

As is explained in more detail below, and contrary to Singular's infringement contentions, Dr. Khatri now recognizes that the MXU multipliers perform exact math—they do not introduce any loss of precision. *See id.* ¶ 162. Thus, in attempting to map the "LPHDR execution unit" element of the patents-in-suit to the accused TPUs, Dr. Khatri relies exclusively on a new, two-stage "LPHDR multiplication operation" infringement theory. Singular never disclosed this theory in either its original or supplemental contentions, even though it had the information upon which it now seeks to rely in advance of the close of fact discovery in September 2021. At no time has Singular sought leave from the Court to amend its infringement contentions to reflect Dr. Khatri's new theory of infringement. Thus, Singular has not complied with the Local Rules and the Scheduling Order, which required Singular to disclose in timely presented infringement contentions the theories upon which it (and thus its expert witness) will seek to rely, and the Court should strike the new infringement theory from Dr. Khatri's expert report.

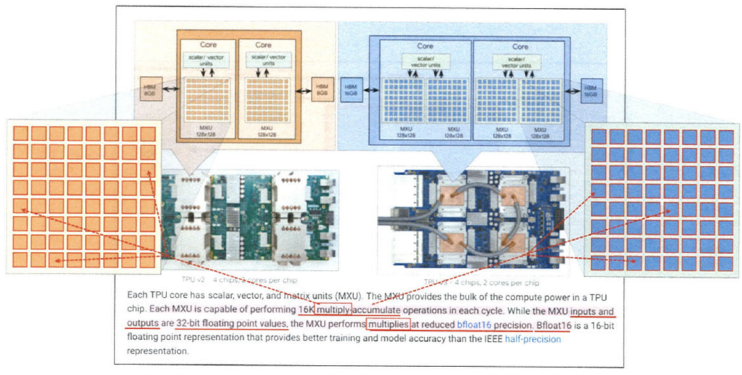
A. Dr. Khatri's report posits a two-stage "LPHDR multiplication operation" that is in fact two separate operations performed by a rounding circuit in the VPU and a multiplier in the MXU.

According to Dr. Khatri's new theory, "the LPHDR execution units of the accused TPUv2 and TPUv3 devices each comprise precision-reducer circuits [in the VPUs] that convert each of the FP32 input signals into low-precision BF16 signals, . . . and a BF16 multiplication circuit [in the MXUs] that multiplies the BF16 signals and produces an output signal." Ex. 1 ¶ 228. Dr. Khatri opines that these separate "circuits" "[c]ollectively" meet the LPHDR execution unit limitation. *Id.* ¶ 140.

Dr. Khatri thus opines that the "LPHDR execution unit" limitation is satisfied by the accused TPUs because they purportedly perform an "LPHDR multiplication operation" in "two stages." *Id.* ¶¶ 128-42. Dr. Khatri acknowledges that he is relying on components in both the VPU and the MXU. *Id.* ¶ 79 (referring to "LPHDR execution units that are included within the VPU/MXU"). Putting aside the merits of whether two separate operations in two separate components can be mapped to the LPHDR execution unit for infringement purposes, this new "two-stage" theory, relying on two components, was not disclosed.

B. Singular's infringement contentions do not disclose an LPHDR execution unit that comprises, in part, "precision-reducer circuits" in the VPU.

In its infringement contentions, Singular did not disclose the purported "LPHDR execution unit" that Dr. Khatri identifies in his report. Instead, in its contentions Singular identified only the multiplier in the MXU as an alleged LPHDR execution unit, pointing only to the multiplication operation performed in the MXU as the operation performed by the LPHDR execution unit:

'156 PATENT	SUPPLEMENTAL INFRINGEMENT EVIDENCE
7. A device comprising: **at least one first low precision high-dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value,** wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input;	 https://cloud.google.com/tpu/docs/system-architecture

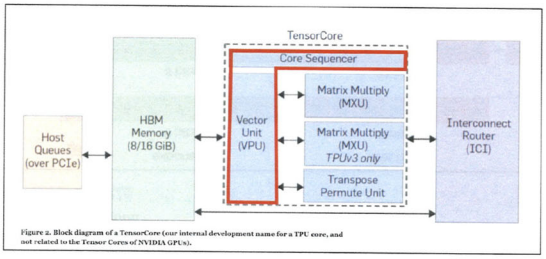
Ex. 3 (2022 08 11 Suppl. Claim Chart for '156 Patent) (Dkt. No. 355-15) at 6 (Singular's color

coding);[1] *see also* Ex. 4 (2022 08 11 Suppl. Claim Chart for '273 Patent) (Dkt. No. 355-16) at 6; Ex. 5 (2020 09 04 Preliminary Claim Chart for '156 Patent) at 6; Ex. 6 (2020 09 04 Preliminary Claim Chart for '273 Patent) at 6. Singular's contentions do not mention a "precision-reducer" or even a "rounding" circuit. Nor do they identify a "two-stage" operation. Rather, the contentions identify *only* the multipliers in the MXU as satisfying the "LPHDR execution unit" limitation.

Likewise, in the portion of its infringement contentions regarding the required error rate, Singular said that for each of the possible valid inputs to the multiplication operation "performed by *the multipliers within the MXU*" it had calculated the result and had concluded that the "MXU multiplier" surpassed the required error rate. Ex. 3 at 9 (emphasis added); *see also* Ex. 4 at 9; Ex. 5 at 9; Ex. 6 at 9. Unlike Dr. Khatri, Singular did not refer to an execution unit in the "VPU/MXU" that performed LPHDR multiplication in a two-stage operation. *See* Ex. 1 ¶ 79.

Indeed, far from disclosing that it was contending that the "precision-reducer circuit" *in the VPU* was part of its accused LPHDR execution unit, Singular asserted that the VPU met a *different* limitation that requires "at least one first computing device adapted to control the at least one first LPHDR execution unit":

[1] Notably, Singular highlighted the phrases about the MXU, but did not highlight the reference in the first sentence to the "vector . . . unit" (i.e., the VPU). *See* Ex. 3 at 6. Similarly, Singular augmented the graphics to add pointers to the individual multipliers within the MXU, but did not identify (i.e., point to) the vector units, much less any "precision-reducer circuits" within the VPU.

'156 PATENT	SUPPLEMENTAL INFRINGEMENT EVIDENCE
7. A device comprising: at least one first low precision high-dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value, wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input; **at least one first computing device adapted to control the operation of the at least one first LPHDR execution unit** wherein the at least one first computing device comprises at least one of a central processing unit (CPU), a graphics processing unit (GPU), a field programmable gate array (FPGA), a microcode-based processor, a hardware sequencer, and a state machine; and, wherein the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.	• "**Each of the cores on a TPU device can execute user computations (XLA ops)** independently." https://cloud.google.com/tpu/docs/system-architecture#pod • "**TPUs use a VLIW architecture to express instruction-level parallelism to the many compute units of a TensorCore. XLA uses standard VLIW compilation techniques** including loop unrolling, instruction scheduling, and software pipelining to keep all compute units busy and to simultaneously move data through the memory hierarchy to feed them." https://cacm.acm.org/magazines/2020/7/245702-a-domain-specific-supercomputer-for-training-deep-neural-networks/fulltext • "The ***Core Sequencer*** fetches *VLIW* (*Very Long Instruction Word*) instructions from the core's on-chip, software-managed Instruction Memory (Imem), executes scalar operations using a 4K 32-bit scalar data memory (Smem) and 32 32-bit scalar registers (Sregs), and forwards vector instructions to the VPU. **The 322-bit VLIW instruction can launch eight operations: two scalar, two vector ALU, vector load and store, and a pair of slots that queue data to and from the matrix multiply and transpose units. The XLA compiler schedules loading Imem via independent overlays of code, as unlike conventional CPUs, there is no instruction cache.**" *Id.* • "The ***Vector Processing Unit (VPU)*** performs vector operations using a large on-chip *vector memory* (*Vmem*) with 32K 128 x 32-bit elements (16MiB), and 32 2D *vector registers* (*Vregs*) that each contain 128 x 8 32-bit elements (4 KiB). **The VPU streams data to and from the MXU through decoupling FIFOs. The VPU collects and distributes data to Vmem** via *data-level parallelism* (2D matrix and vector functional units) and *instruction-level parallelism* (8 operations per instruction)." *Id.*  Figure 2. Block diagram of a TensorCore (our internal development name for a TPU core, and not related to the Tensor Cores of NVIDIA GPUs). *Id.* *See also* GOOG-SING-SC-62-227, 258-267, 269-289, 346-354, 356-373, 445-448.

Ex. 3 at 10 (Singular's color coding). Thus, instead of identifying the "precision-reducer circuits" in the VPU as being components of its accused LPHDR execution units (as Dr. Khatri now does), Singular's infringement contentions identified the VPU as a computing device that *controls* the LPHDR execution unit. *Id.* Further reflecting the changed infringement theory, Dr. Khatri's new opinion omits mapping the "device adapted to control" to the VPU.[2] *See* Ex. 1 ¶¶ 209-17.

Dr. Khatri's departure from Singular's contentions is underscored by his mapping of the claim limitation requiring that the number of LPHDR execution units in the accused device "exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at

[2] As indicated in the figure above, Singular's infringement contentions highlighted the Core Sequencer and the VPU as meeting the "control" limitation. Dr. Khatri's report, however, relies only on the Core Sequencer for this limitation and does not cite the VPU. *See* Ex. 1 ¶¶ 209-17.

least 32 bits wide." *See* Ex. 1 ¶ 222. Singular previously mapped this limitation to the multipliers in the MXU, consistent with its prior contention that those multipliers were the LPHDR execution units. *See* Ex. 3 at 12; Ex. 4 at 10. Dr. Khatri now identifies the "precision-reducer circuits" as part of what he counts in order to meet the exceeds limitation. But by his own count, the TPUs have fewer "precision-reducer circuits" than units that perform at least the operation of multiplication at 32 bits. *See* Ex. 1 ¶¶ 232, 234. Thus, to meet the "exceeds" limitations, Dr. Khatri counts each rounding circuit of his purported LPHDR execution units multiple times, claiming that it is "routine practice in all areas of engineering" to use computing "cycles" to "count" a piece of hardware multiple times in defining what constitutes an arithmetic unit. *See id.* ¶ 233. Setting aside the merits of that position,[3] it was not disclosed in Singular's contentions.

Because Singular did not previously disclose the infringement theories in Dr. Khatri's expert report, they are improper and should be stricken, particularly at this late stage of the case. *See, e.g., ViaTech*, 2021 WL 663057, at *5 ("I do not understand how Plaintiff's experienced lawyers could have thought that springing clearly new theories on a defendant in opening expert reports was in compliance with the scheduling order, the Rules, or expected standards of practice."); *KlausTech*, 2018 WL 5109383, at *3 (noting that "expert reports cannot go beyond the bounds of the disclosed infringement theories and introduce new theories not disclosed in the contentions").

[3] Dr. Khatri's assertion of what is engineering practice, whether true or not, has no bearing on the patent law question of whether, in mapping the elements of an apparatus claim requiring numerosity, one could legitimately count the same piece of hardware multiple times to meet the numerosity requirement. To the extent Singular is permitted to pursue this line of argument despite it not being disclosed in Singular's contentions, Google expects to challenge this position in the context of summary judgment and/or *Daubert* motion practice.

C. Dr. Khatri's new theory of infringement is not based on new information but is instead a tactical shift attempting to avoid the problems with the infringement theory in Singular's contentions.

Dr. Khatri's report does not rely on new information that Singular recently received or did not have access to prior to the close of fact discovery in September 2021. Even the source code that Dr. Khatri cites in his report was made available to Singular on a source code computer prior to the close of fact discovery. *See* Ex. 1 ¶¶ 143-45. Indeed, Dr. Khatri's "Materials Considered" does not identify any materials that were not available to Singular prior to the close of fact discovery. *See* Ex. 7.

Dr. Khatri's new theory reflects a tactical shift necessitated in part by his concession that the MXU multipliers perform exact math. Dr. Khatri admits that "[t]he BF16 multiplication in the second stage of the multiplication operation [i.e. in the MXU multiplier] multiplies the two rounded BF16 signals *without further loss of precision.*" Ex. 1 ¶ 162 (emphasis added). Given Dr. Khatri's concession that the MXU multipliers perform exact math, these multipliers, which Singular had previously contended met the LPHDR execution unit element, cannot meet the claimed error requirements of the LPHDR execution unit.[4]

D. The Court should strike Singular's new, replacement infringement theory.

The Court should strike the new infringement theory in Dr. Khatri's expert report based on Singular's failure to comply with the Local Rules and the Scheduling Order prior to introducing new theories of infringement. *ViaTech*, 2021 WL 663057, at *5; *KlausTech*, 2018 WL 5109383, at *3. Indeed, when it moved to supplement its infringement contentions Singular expressly denied that it was disclosing a new infringement theory, and represented that it had added nothing beyond

[4] Dr. Khatri's theory suffers from its own problems on the merits, which Google will address if necessary at an appropriate time. But because Singular did not timely disclose Dr. Khatri's two-stage "LPHDR multiplication operation," the theory should be stricken from Dr. Khatri's expert report.

source code citations. It accordingly made no effort to demonstrate good cause to replace its infringement theory, let alone to establish that Google would not be prejudiced by such a change.

Regardless, as the Court has recognized, "courts generally disfavor amendments to add new theories of infringement, particularly late in the litigation or near or after the close of discovery, as unduly prejudicial." *Philips*, 2021 WL 5417103, at *5 (collecting cases). Fact discovery ended in September 2021 and trial is set for September 2023. *See* Dkt. Nos. 70, 212, 302, 403. Google has formulated its defenses in this action, including its invalidity defenses, based on the infringement theories that Singular disclosed in discovery. It is far too late for Singular to be changing its position now.

V. CONCLUSION

For the foregoing reasons, Google respectfully requests that the Court strike the portions of Dr. Khatri's expert report that rely on a two-stage "LPHDR multiplication operation," including but not limited to paragraphs 128-133, 140-142, 144-146, 148, 158-174, 178-210, 212, 222-223, 228-233, 235, and 237-238, and also the exhibits to his report that are cited only in stricken portions of the report, including but not limited to Exhibits D-G.

Respectfully submitted,

Dated: January 18, 2023

By: */s/ Nathan R. Speed*

Gregory F. Corbett (BBO #646394)
gcorbett@wolfgreenfield.com
Nathan R. Speed (BBO # 670249)
nspeed@wolfgreenfield.com
Elizabeth A. DiMarco (BBO #681921)
edimarco@wolfgreenfield.com
Anant K. Saraswat (BBO #676048)
asaraswat@wolfgreenfield.com
WOLF, GREENFIELD & SACKS, P.C.
600 Atlantic Avenue
Boston, MA 02210
Telephone: (617) 646-8000
Fax: (617) 646-8646

Robert Van Nest (admitted *pro hac vice*)
rvannest@keker.com
Michelle Ybarra (admitted *pro hac vice*)
mybarra@keker.com
Andrew Bruns (admitted *pro hac vice*)
abruns@keker.com
Vishesh Narayen (admitted *pro hac vice*)
vnarayen@keker.com
Christopher S. Sun (admitted *pro hac vice*)
csun@keker.com
Anna Porto (admitted *pro hac vice*)
aporto@keker.com
Deeva Shah (admitted *pro hac vice*)
dshah@keker.com
Stephanie J. Goldberg (admitted *pro hac vice*)
sgoldberg@keker.com
KEKER, VAN NEST & PETERS LLP
633 Battery Street
San Francisco, CA 94111-1809
Telephone: (415) 391-5400

Michael S. Kwun (admitted *pro hac vice*)
mkwun@kblfirm.com
Asim M. Bhansali (admitted *pro hac vice*)
abhansali@kblfirm.com
KWUN BHANSALI LAZARUS LLP
555 Montgomery Street, Suite 750
San Francisco, CA 94111
Telephone: (415) 630-2350
Fax: (415) 367-1539

Matthias A. Kamber (admitted *pro hac vice*)
matthiaskamber@paulhastings.com
PAUL HASTINGS LLP
101 California Street, 48th Floor
San Francisco, CA 94111
Telephone: (415) 856-7000
Fax: (415) 856-7100

Counsel for Defendant Google LLC

CERTIFICATE OF SERVICE

I certify that this document is being filed through the Court's electronic filing system, which serves counsel for other parties who are registered participants as identified on the Notice of Electronic Filing (NEF). Any counsel for other parties who are not registered participants are being served by first class mail on the date of electronic filing.

/s/ Nathan R. Speed
Nathan R. Speed